Lung Cancer Screening with Submillisievert Chest CT: Potential Pitfalls of Pulmonary Findings in Different Readers with Various Experience Levels

Martini, Katharina ; Ottilinger, Thorsten ; Serrallach, Bettina ; Markart, Stefan ; Glaser-Gallion, Nicola ; Blüthgen, Christian ; Leschka, Sebastian ; Bauer, Ralf W ; Wildermuth, Simon ; Messerli, Michael

Abstract: Purpose To assess the interreader variability of submillisievert CT for lung cancer screening in radiologists with various experience levels. Method Six radiologists with different degrees of clinical experience in radiology (range, 1-15 years), rated 100 submillisievert CT chest studies as either negative screening finding (no nodules, benign nodules, nodules <5 mm), indeterminate finding (nodules 5-10 mm), positive finding (nodules >10 mm). Each radiologist interpreted scans randomly ordered and reading time was recorded. Interobserver agreement was assessed with ak statistic. Reasons for differences in nodule classification were analysed on a case-by-case basis. Reading time was correlated with reader experience using Pearson correlation (r). Results The overall interobserver agreement between all readers was moderate ($k = 0.454$; $p < 0.001$). In 57 patients, all radiologists agreed on the differentiation of negative and indeterminate/positive finding. In 64 cases disagreement between readers led to different nodule classification. In 8 cases some readers rated the nodule as benign, whereas others scored the case as positive. Overall, disagreement in nodule classification was mostly due to failure in identification of target lesion ($n = 40$), different lesion measurement ($n = 44$) or different classification ($n = 26$). Mean overall reading time per scan was of 2 min 2 s (range: 7s-7 min 45 s) and correlated with reader-experience ($r = -0.824$). Conclusions Our study showed substantial interobserver variability for the detection and classification of pulmonary nodules in submillisievert CT. This highlights the importance for careful standardisation of screening programs with the objective of harmonizing efforts of involved radiologists across different institutions by defining and assuring quality standards.

DOI: https://doi.org/10.1016/j.ejrad.2019.108720

Lung Cancer Screening with Submillisievert Chest CT: Potential Pitfalls of Pulmonary Findings in Different Readers with Various Experience Levels. European Journal of Radiology, 121:108720.
DOI: https://doi.org/10.1016/j.ejrad.2019.108720

Journal Pre-proof

Lung Cancer Screening with Submillisievert Chest CT: Potential Pitfalls of Pulmonary Findings in Different Readers with Various Experience Levels

Katharina Martini, Thorsten Ottilinger, Bettina Serrallach, Stefan Markart, Nicola Glaser-Gallion, Christian Blüthgen, Sebastian Leschka, Ralf W. Bauer, Simon Wildermuth, Michael Messerli

PII: S0720-048X(19)30370-5

DOI: https://doi.org/10.1016/j.ejrad.2019.108720

Reference: EURR 108720

To appear in: *European Journal of Radiology*

Received Date: 31 December 2018

Revised Date: 3 October 2019

Accepted Date: 21 October 2019

Please cite this article as: Martini K, Ottilinger T, Serrallach B, Markart S, Glaser-Gallion N, Blüthgen C, Leschka S, Bauer RW, Wildermuth S, Messerli M, Lung Cancer Screening with Submillisievert Chest CT: Potential Pitfalls of Pulmonary Findings in Different Readers with Various Experience Levels, *European Journal of Radiology* (2019), doi: https://doi.org/10.1016/j.ejrad.2019.108720

Lung Cancer Screening with Submillisievert Chest CT: Potential Pitfalls of Pulmonary Findings in Different Readers with Various Experience Levels

Type of manuscript: Original research

[a,b]Katharina Martini, MD, [c]Thorsten Ottilinger, MD, [c]Bettina Serrallach, MD, [c]Stefan Markart, MD, [c]Nicola Glaser-Gallion, MD, [a,b]Christian Blüthgen, MD, [b,c]Sebastian Leschka, MD, [c,d]Ralf W. Bauer, MD, EBCR, [c]Simon Wildermuth, MD, [a]Michael Messerli, MD

[a]Department of Nuclear Medicine, University Hospital Zurich, University Zurich, Switzerland
[b]Institute of Diagnostic and Interventional Radiology, University Hospital Zurich, University Zurich, Switzerland
[c]Division of Radiology and Nuclear Medicine, Cantonal Hospital St. Gallen, Switzerland
[d]RNS Gemeinschaftspraxis GbR, Wiesbaden, Germany

Corresponding author: Dr. Michael Messerli
Department of Nuclear Medicine
University Hospital Zurich
Ramistrasse, Zurich, Switzerland.
Phone: +41-44-255 18 18
Fax: +41-44-255 44 14
E-Mail: michael.messerli@usz.ch

Highlights:

- Variability of lung nodule classification in submillisievert CT was assessed
- Overall interobserver agreement between all readers was moderate (k = 0.454)
- Major reason for disagreement on nodule classification was lesion measurement
- Lung cancer screening with submillisievert CT needs careful standardisation

Abstract

Purpose: To assess the interreader variability of submillisievert CT for lung cancer screening in radiologists with various experience levels.

Method: Six radiologists with different degrees of clinical experience in radiology (range, 1-15 years), rated 100 submillisievert CT chest studies as either negative screening finding (no nodules, benign nodules, nodules <5 mm), indeterminate finding (nodules 5-10 mm), positive finding (nodules >10 mm). Each radiologist interpreted scans randomly ordered and reading time was recorded. Interobserver agreement was assessed with a *k* statistic. Reasons for differences in nodule classification were

analysed on a case-by-case basis. Reading time was correlated with reader experience using Pearson correlation (*r*).

Results: The overall interobserver agreement between all readers was moderate (*k*=0.454; *p*<0.001). In 57 patients, all radiologists agreed on the differentiation of negative and indeterminate/positive finding. In 64 cases disagreement between readers led to different nodule classification. In 8 cases some readers rated the nodule as benign, whereas others scored the case as positive. Overall, disagreement in nodule classification was mostly due to failure in identification of target lesion (*n*=40), different lesion measurement (*n*=44) or different classification (*n*=26). Mean overall reading time per scan was of 2 min 2 s (range: 7s-7min 45s) and correlated with reader-experience (r =-0.824).

Conclusions: Our study showed substantial interobserver variability for the detection and classification of pulmonary nodules in submillisievert CT. This highlights the importance for careful standardisation of screening programs with the objective of harmonizing efforts of involved radiologists across different institutions by defining and assuring quality standards.

Keywords: CT; Low dose; Lung cancer; Screening; Radiation dosage; Iterative reconstruction

Abbrevations:
ADMIRE advanced modelled iterative reconstruction
CT computed tomography
NLST National Lung Screening Trial
NELSON The Dutch-Belgian Lung Cancer Screening Trial

1. INTRODUCTION

Lung cancer is the leading cause of cancer deaths in Europe and around the world [1]. Over 50% of newly diagnosed lung cancer patients are former, not current, smokers [2] and the combination of poor outlook, lag time and large population “at risk” means that lung cancer will remain a significant disease burden over the coming years

[3]. As the population at risk is relatively well defined and early stage disease is potentially curable, lung cancer outcomes may be improved by screening. The main target of lung cancer screening programs with computed tomography (CT) is an early detection and treatment of potentially malignant pulmonary lesions to improve clinical outcomes, as relative survival rates without recurrence of non-small cell lung cancer is up to 80% when detected at a locally confined stage [4; 5]. With positive results from the National Lung Screening Trial (NLST) that suggested a reduced lung cancer mortality among people at high risk [6] the interest in screening increased over the past decade. Large multicentre studies in Europe such as the Dutch-Belgian randomised lung cancer multi-slice CT screening trial (NELSON) also reported favourable results but are currently ongoing [7; 8], yet different societies already recommend lung cancer screening [9; 10].

In every lung cancer screening program, subjects undergo multiple screening rounds within a certain time period and in order to keep the cumulative radiation exposure reasonable, so-called “low dose” CT protocols are established. Various technical innovations (e.g., tin filtration, iterative image reconstruction) in the past years allowed for a steady and even bigger reduction of applied radiation dose beyond standard “low dose” protocols, permitting submillisievert lung CT at the dose of a two-view chest X-ray examination [11]. Recent studies have already reported a high diagnostic accuracy of this technique for the detection and assessment of different pathologies in the lung, mainly focussing on pulmonary nodule detection [11-13].

Beyond sole detection, the correct classification of CT screening studies as either negative (i.e., no or only benign nodules present) or positive (i.e., pulmonary nodules present) among different radiologists is crucial for an effective screening programme and has a major effect on patient management. Reproducible and accurate classification by radiologists ideally across different experience levels is

desired and has to be guaranteed, especially when extremely lowering the utilized radiation dose levels.

Accordingly, the aim of our study was to assess interreader agreement of chest X-ray-equivalent dose CT for lung cancer screening in radiologists with various levels of experience and to identify and report potential pitfalls of uniform image interpretation.

2. MATERIALS AND METHODS

2.1. Overview and study design

The patients included in this study are part of a prospective study at our institution on reduced radiation dose CT (clinicaltrials.gov Identifier *blinded for review*). This is an investigator-initiated study and no funding was received by any funding agencies in the public, commercial, or not-for-profit sectors. The institutional ethics committee approved the study. All patients included in this study gave written informed consent for a CT with submillisievert dose additionally to a clinically indicated standard dose CT. The study was conducted in compliance with ICH-GCP-rules and the declaration of Helsinki.

2.2. Patients

One hundred patients that were referred to our division for a clinically indicated CT between February and June 2015 were prospectively included in the study and scanned with an additional submillisievert CT of the chest in the same session. Inclusion criteria were a clinically indicated chest CT with or without contrast media for various indications (e.g., follow up of pulmonary nodule, suspicion of pulmonary embolism, work up of suspicious lesion in chest X-ray). Exclusion criteria for being included in the study was (a) pregnancy and/or (b) age < 18 years. The study group has been partly shared and is described more in detail in a previous publication [14].

2.3. Reduced dose (submillisievert) CT protocol

All scans were performed using a third-generation dual-source 192 section CT scanner (SOMATOM Force; Siemens Healthineers, Forchheim Germany) [15]. Patients were scanned according a CT protocol that was previously described in detail [14]. A similar scan protocol assessing pulmonary nodule detection rate was used in a phantom study by Gordic et. al [11].

Dose-length-product (DLP) and volume CT dose index $CTDI_{vol}$ were retrieved from the electronically logged patient protocol. The effective radiation dose was then calculated by multiplying the DLP with the conversion coefficient k of 0.014 mSv/mGycm [16]. We further assessed the effective diameters of the chest for each subject and calculated the size-specific dose estimates (SSDE) applying the size-specific conversion factor f_{size} from the AAPM Report 204 ($SSDE = f_{size} \times CDTI_{vol}$) [17].

2.4. Data reconstruction

CT images were reconstructed with advanced modelled iterative reconstruction (ADMIRE; Siemens Healthineers) as described in detail before [11] at a strength level of 3 using a slice thickness of 2 mm with an increment of 1.6 mm and an edge-enhancing convolution kernel (Br64). The reconstructed field-of-view (FoV) was in general 400 mm and in patients with a body-mass-index of >30 kg/m^2 480 mm. The image matrix size was 512 x 512 pixels. Image analyses were performed on a picture archiving and communication system workstation using Impax EE (Version R20 XV SU2; Agfa Healthcare N.V., Mortsel, Belgium) and a standard high-definition liquid crystal display monitor for clinical reporting (BARCO; Medical Imaging Systems, Kortrijk, Belgium).

2.5. CT data analysis

Submillisievert CT studies of 100 patients were assigned to the readers in random orders with a random number generator (http://stattrek.com/Tables/Random.aspx). Radiologists completed an evaluation spread sheet assessing CT series (lung window CT setting; window level of -600 HU and a width of 1200 HU) for the presence of pulmonary nodules. As pulmonary nodules the readers classified circumscribed opacities in the pulmonary interstitium with increased density – i.e., a pulmonary nodule according the Fleischner Society [18].

Subpleural nodules and nodules along the fissures were not included. Readers did not differ between solid, part-solid or subsolid nodules.

If a nodule was present, they had to measure axial nodule diameter (i.e., longest diameter) and to classify studies as follows: 1) no nodule present (N1); clearly benign nodule containing calcification (N2/C); clearly benign nodule containing fat (N2/F); any other nodule < 5 mm (N3); nodules between 5 – 10 mm (I); nodules > 10 mm (P). These findings were further subdivided in “negative findings” (N1, N2/C, N2/F, N3) where no further action is needed and indeterminate findings (I) as well as positive findings (P) were further work up is required, according **Table 1** and **Figure 1**. If multiple pulmonary nodules were present, the most suspicious nodule was defined as “target lesion” and used as guide for management according to the guidelines of the Fleischner Society [19]. Readers were allowed to change window settings and to magnify images at will. They were blinded to the other radiologists’ findings. The time it took for the complete evaluation of the study was recorded. The rationale for choosing the aforementioned somewhat simplified “nodule classification” system was as follows: Currently, neither in our department nor in any institution in our country a lung cancer screening is performed, which is why the radiologists are not familiar with any lung cancer screening guideline (e.g., American College of Radiology, ACR; nor British Thoracic Society, BTS). A table linking our classification system to the ACR/Lung-RADS system (source: https://www.acr.org/Clinical-Resources/Reporting-and-Data-Systems/Lung-Rads) can be found in **Figure 2**.

2.7. Statistical analysis

Continuous variables are expressed as mean ± SD, and categorical variables are expressed as frequencies or percentages. Cohen’s Kappa (*k*) was used to assess interreader agreement for nodule classification. The resulting *k*-value were stratified

qualitatively as follows: slight agreement, 0.01–0.20; fair agreement, 0.21–0.40; moderate agreement, 0.41–0.60; substantial agreement, 0.61–0.80; excellent agreement, 0.81–0.99 [20]. The absolute number of agreements/disagreement was calculated considering also different classifications per case. Further, reading time was correlated to reading experience using Pearson-correlation (*r*). All statistical analyses were conducted using commercially available software (SPSS, release 22.0; SPSS, Chicago, IL, USA). A two-tailed *p*-value of 0.05 was considered to indicate statistical significance.

3. RESULTS

One hundred patients (36 female; median age 63 years; range 18 - 79 years) were included in the study and scanned with a submillisievert CT of the chest.

3.1. Radiation dose values of the study protocol

The median effective radiation dose of was 0.13 mSv (range, 0.11 – 0.16 mSv). The median DLP was 9.5 mGy-cm (range, 7.5 – 11.6 mGy-cm) and $CTDI_{vol}$ was 0.24 mGy. The median effective diameter was 30 cm (range, 22 – 39 cm) and the median SSDE was 0.30 mGy (range, 0.21 – 0.40 mGy).

3.2. Interobserver agreement

The overall interobserver agreement between all readers was moderate (k = 0.454; $p < 0.001$). When evaluating interobserver agreement between experienced readers, agreement levels were lower ($k = 0.404$, $p < 0.001$) compared to their younger colleagues ($k = 0.544$; $p < 0.001$). For interreader agreement in nodule classification, please refer to **Table 2**. Nodule classification per reader is presented in **Table 3** and **Figure 3**.

3.3. Positive/indeterminate vs. negative screening results

Overall, in the differentiation of positive/indeterminate finding (i.e., further work-up needed) and negative (i.e., no further action required) in 21 patients all radiologists agreed that the findings are indeterminate or positive and need further work-up and in 36 patients all radiologists agreed that the patient has a negative CT scan, **Figure 4**.

In the 22 cases where at least one reader rated the scan as positive (P) not all readers agreed and rated the scan as either indeterminate or negative: In 13 (59%) cases other readers assigned the scan as indeterminate, in 8 (36%) cases there was

at least one reader who rated the scan as negative (i.e., N1-3) and in 8 (36%) cases at least one reader rated that there was no nodule present (i.e., N1), see **Figure 4A**.

In 43 cases where at least one reader rated the scan as positive (P) or indeterminate (I), meaning that the patient needs further work-up, not all readers agreed and assigned the scan as negative (i.e., N1-3): In 37 (86%) cases other readers assigned the scan as benign (i.e., N2) and in 21 (49%) cases at least one reader rated that there was no nodule present (i.e., N1), see **Figure 4B**.

There was substantial variation in the total number of nodules recorded, with some readers identifying more than twice as many patients that have a nodule than others: In 45 patients at least one reader assigned “no pulmonary nodule found” (i.e., N1) even though other readers detected nodules. In these cases, 40 (89%) scans of the detected nodules were rated as negative (i.e., only benign and/or nodules < 5 mm; N2-3) by the other readers and in 21 (47%) cases at least one reader rated the findings as indeterminate (n = 16; 36%) or positive (n = 8; 18%) (**Figure 4**).

3.4. Lesion size agreement

In 23 cases variation of nodule measurements lead to different nodule classification. In 10 cases different nodule measurement and resulting different classification (B vs. I or P) affected patient work-up. The higher the nodule diameter, the higher were measurement differences among the readers resulting in a higher standard deviation and higher mean range of measurements, **Table 4**.

3.5. Reasons for disagreement among readers

A differentiated case-by-case evaluation of ratings of the different readers showed that the misclassification of the nodules was mostly due to failure of identification of the target lesion (n = 40), especially in cases where the disagreement was benign finding (N1-N3) vs. positive (P) finding. Differences of size measurement

accounted for $n = 44$ misclassified cases and affected especially the discrimination of benign from indeterminate findings. Another frequent pitfall was failure of detection of calcium ($n = 24$) or fat ($n = 2$). Another common reason for CT mis-classification were missed nodules: Overall, in 29% of cases reason for different CT classification were missed nodules – in 27% of cases missed nodules led to different patient management, **Table 5** and **Figure 5**. The number of missed lesions did not significantly differ between experienced and unexperienced radiologists and no specific location (lung lobe) was identified where nodules were missed more often ($p > 0.05$).

3.6. Reading time

Mean overall reading time was 2 minutes and 2 seconds per scan, whereas minimal reading time was 7 seconds and maximum time was 7 minutes and 45 seconds.

The three more experienced radiologists were significantly faster in reading CT scans compared to the three less experienced radiologists (mean 1 minute, range 7 seconds to 2 minutes 39 seconds vs. mean 3 minutes 3 seconds, range 14 seconds to 7 minutes 45 seconds; $p < 0.001$). Mean overall reading time per scan correlated with reader-experience ($r = -0.867$, $p = 0.04$), **Figure 6**.

4. DISCUSSION

In our study we sought to assess the interreader agreement of chest X-ray dose-equivalent (submillisievert) CT for correct nodule detection and classification and to identify and report potential pitfalls of pulmonary findings among radiologists with various levels of experience. While previous studies mainly focussed on nodule detectability [12; 13], this is, to our knowledge the first study to assess interreader agreement of pulmonary findings at such a low (submillisievert) radiation dose.

Recently, multiple lung cancer screening studies using CT are ongoing in different countries [21; 22]. The National Lung Screening Trial in the United States [21] and the NELSON Trial in Holland/Belgium [8] have both presented encouraging results. In order to guarantee a high-quality expert chest CT report, an evaluation of the eventual pitfalls in screening CT by radiologists and education is needed.

Our results indicate that there is a substantial interobserver variability for the detection and classification of pulmonary nodules, which may have impact on patient management.

The study showed that in more than 40% of patients at least one radiologist did not agree with the reading of the others. This was partly due to a high interreader variability in the sub-classification of benign lesions and the notable number of benign lesions, which were not detected by a group of readers. The risk for malignancy in small nodules with a diameter of < 5 mm is very low with less than 1% [23]. This is also reflected in the 2015 published guidelines for the investigation and management of pulmonary nodules from the BTS [24], in which no nodule follow-up for people with nodules < 5 mm in maximum diameter is recommended. Similarly, the Fleischner Society states that in patients which are eligible for lung cancer screening (i.e., so called high risk patients such as smokers) nodules < 6 mm do not require routine follow up [19]. Failure in the identification or discrimination of benign lesions (N1 to N3) can

therefore be ignored since they do not have impact on patient management. Similar to Gierada et al. [25] we observed a higher agreement for classification of cases as positive or indeterminate findings than of negative (i.e., benign) lesions.

Nevertheless, there was also a substantial interreader variability in the classification of positive (P) and potential malignant lesions; > 10 mm, and indeterminate (I) lesions which may urge follow-up or initiation of further workup and may therefore have impact on patient management (e.g., by a delay of diagnosis or by unnecessary invasive work-up).

Failure in nodule classification is not a new finding and was already reported by other authors [25-27]. For example, Ridge et al. [26] reported a high interreader variability in the differentiation of solid and subsolid pulmonary nodules (i.e., nodule classification). Gierada et al. [25] evaluated agreement among radiologists on the interpretation of pulmonary findings in screening examination for lung cancer and reported only a moderate to substantial interobserver agreement (k = 0.58) contemplating a potential for considerable improvement. Similar to our findings this study reported a wide range in the total number of lesions detected with some radiologists identifying several times as many non-calcified nodules than others [25].

In our study, relevant interreader differences were mainly due to variation of the identified target lesion, differences in classification of target lesions, discrimination between indeterminate (I) and positive lesions (P) especially in borderline sizes; the latter mainly due to variation of nodule measurements. An additional pitfall was that small lesions were often overlooked, however as already stated before small/benign lesions do not have to be followed-up and therefore the missed finding potentially does not have any impact in further patient management. Interestingly, the number of missed lesions did not differ significantly between experienced and unexperienced radiologists.

With regard to a screening setting, this highlights the importance for a rigorous training of involved radiologists and underlines the need for quality assurance systems with sufficient quality control mechanisms to provide high diagnostic accuracy for the screened population. Further, regular multidisciplinary review meetings should be held to discuss cases and decide on management as it is already recommended for breast cancer screening by the European Breast Cancer Network (EBCN) [28]. This may prove beneficial for feedback purposes of involved radiologists as well as for providing an optimized mechanism for refining individual case management decisions. Later in the course of a screening, the review of interval cancers by involved readers, as part of an organised process may also serve as an excellent feedback and educational mechanism [28]. As with regard to lung cancer screening, the ACR issued a quality assurance tool designed to standardize screening CT reporting and management recommendations, in order to reduce confusion in lung cancer screening imaging interpretations, and to facilitate outcome monitoring (i.e., so called Lung-RADS reporting system). Since we used slightly different size threshold cut-offs in our study no direct transformation of our findings is possible to centres using ACR or BTS guidelines. We have therefore amended our readout spreadsheet as supplementary material allowing individual assessment of respective variation relevant for different institutions.

As stated before, differences in lesion measurement substantially contributed to disagreement among the readers. Variance in lesion measurement was a major cause of nodule mis-classification (28% overall – and 30% in cases were different measurement let to different patient management). This goes in line with earlier studies [29; 30], which have shown considerable variation in two-dimensional lung nodule size measurements. Nodule size at baseline scan correlates with lung cancer risk [31], has

impact on patient management and thus the measurement issue should be addressed in the near future e.g. by automated volume measurement.

Screening results and subsequent actions taken depend lastly on the sensitivity and interpretation skills of the radiologist. With some lesions, classification of screening findings as positive or negative is not a straightforward task and may depend on individual judgement and experience of the radiologist [25]. Interestingly, we observed an even lower interreader agreement among the experienced readers (k = 0.404) as compared to their less experienced colleagues (k = 0.544).

Aside the training of radiologist, development of evidence-based nodule characterization criteria [32] and automated nodule characterization algorithms [33] (e.g., including texture analysis) may also help increase agreement in nodule classification.

With regard to radiation dose, the “low dose” protocol used in the NLST had a mean dose of 1.5 mSv per scan [6], being substantially higher compared to 0.13 mSv achieved in our study. There has, however, been some debate about how CT-related radiation dose values should be reported [34]. Bankier et al. argued that terms like “low dose” and “ultralow dose” are substantially limited by their relativistic foundation, and therefore recommended not to use them. We have therefore used a more descriptive term (i.e., submillisievert) for our study protocol but more importantly have also reported more meaningful radiation dose parameters (DLP, $CTDI_{vol}$, effective diameter, SSDE).

Our study has some limitations. First, the screening setting was only simulated evaluating CTs obtained for other clinical questions. We are aware, that inclusion criteria for lung cancer screening might be different, however, the discrepancy of pulmonary findings (i.e. nodule measurement/nodule classification) should not be affected by different inclusion criteria. Second, we have to acknowledge the inherent

limitations of in vivo studies on pulmonary nodule detection performance because no gold standard is available, however as our study focuses on the effect of interobserver variability and its impact on management we feel that no reference standard is required. Third, we did not reconstruct CT images with filtered back projection or varying strength levels of ADMIRE to test for its potential influence on nodule classification. Fourth, our readers were not trained for CT lung cancer screening. Further studies may assess whether a pre- and/or interim-screening training of radiologists can improve interreader agreement among different readers. Fifth, it may be regarded as a limitation that we used a somewhat simplified nodule classification system, given that our readers are not currently using any established guideline due to a lack of lung cancer screening in our country. However, we have added the readout spreadsheet from our study as supplementary material and further encourage future studies to assess nodule classification variability using other screening guidelines. Sixth, we did not use semi-/automated volume assessment and test its impact on measurement agreement as well as accuracy of lesion classification. Finally, we did not evaluate nodule structure (e.g., margins, density, solid/subsolid type) and test whether these differences may affect nodule classification. This should be elucidated by future studies.

5. CONCLUSION

In conclusion, our study showed a substantial interobserver variability for the classification of submillisievert CT for lung cancer screening, notably also among experienced radiologists. This highlights the importance for careful standardisation of screening programs with the objective of harmonizing efforts of involved radiologists across different institutions by defining and assuring quality standards. In addition, further studies are warranted to assess whether semi-automated detection, measurement and classification of pulmonary lesions may reduce the disagreement amongst readers.

Conflicts of interest: Prof. Dr. Ralf Bauer is on the speakers' bureau of Siemens Healthineers. Otherwise, we disclose no financial support or author involvement with organizations with financial interest in the subject matter. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Acknowledgements: Dr. Michael Messerli has received a research grant from the Iten-Kohaut Foundation, Switzerland.
We thank Elisabeth Wismer and her team for their excellent technical support.

REFERENCES

1 B S, CP W (2014) World Cancer Report 2014International Agency for Research on Cancer, pp 11

2 Yang P, Allen MS, Aubry MC et al (2005) Clinical features of 5,628 primary lung cancer patients: experience at Mayo Clinic from 1997 to 2003. Chest 128:452-462

3 Marshall HM, Bowman RV, Ayres J et al (2015) Lung cancer screening feasibility in Australia. Eur Respir J 45:1734-1737

4 Carr SR, Schuchert MJ, Pennathur A et al (2012) Impact of tumor size on outcomes after anatomic lung resection for stage 1A non-small cell lung cancer based on the current staging system. J Thorac Cardiovasc Surg 143:390-397

5 Cerfolio RJ, Bryant AS (2009) Survival of patients with true pathologic stage I non-small cell lung cancer. Ann Thorac Surg 88:917-922; discussion 922-913

6 National Lung Screening Trial Research T, Aberle DR, Adams AM et al (2011) Reduced lung-cancer mortality with low-dose computed tomographic screening. N Engl J Med 365:395-409

7 Field JK, van Klaveren R, Pedersen JH et al (2013) European randomized lung cancer screening trials: Post NLST. J Surg Oncol 108:280-286

8 Ru Zhao Y, Xie X, de Koning HJ, Mali WP, Vliegenthart R, Oudkerk M (2011) NELSON lung cancer screening study. Cancer Imaging 11 Spec No A:S79-84

9 Kauczor HU, Bonomo L, Gaga M et al (2015) ESR/ERS white paper on lung cancer screening. Eur Respir J 46:28-39

10 Wender R, Fontham ET, Barrera E, Jr. et al (2013) American Cancer Society lung cancer screening guidelines. CA Cancer J Clin 63:107-117

11 Gordic S, Morsbach F, Schmidt B et al (2014) Ultralow-Dose Chest Computed Tomography for Pulmonary Nodule Detection First Performance Evaluation of Single Energy Scanning With Spectral Shaping. Investigative Radiology 49:465-473

12 Messerli M, Kluckert T, Knitel M et al (2016) Computer-aided detection (CAD) of solid pulmonary nodules in chest x-ray equivalent ultralow dose chest CT - first in-vivo results at dose levels of 0.13mSv. European Journal of Radiology 85:2217-2224

13 Messerli M, Kluckert T, Knitel M et al (2017) Ultralow dose CT for pulmonary nodule detection with chest x-ray equivalent dose - a prospective intra-individual comparative study. Eur Radiol. 10.1007/s00330-017-4739-6

14 *blinded* (*blinded for review*)

15 Morsbach F, Desbiolles L, Plass A et al (2013) Stenosis quantification in coronary CT angiography: impact of an integrated circuit detector with iterative reconstruction. Invest Radiol 48:32-40

16 Task Group on Control of Radiation Dose in Computed T (2000) Managing patient dose in computed tomography. A report of the International Commission on Radiological Protection. Ann ICRP 30:7-45

17 Boone JM, Strauss KJ, Cody D (2011) Size-Specific Dose Estimates (SSDE) in Pediatric and Adult Body CT Examinations. Report of AAPM Task Group 204

18 Hansell DM, Bankier AA, MacMahon H, McLoud TC, Muller NL, Remy J (2008) Fleischner Society: glossary of terms for thoracic imaging. Radiology 246:697-722

19 MacMahon H, Naidich DP, Goo JM et al (2017) Guidelines for Management of Incidental Pulmonary Nodules Detected on CT Images: From the Fleischner Society 2017. Radiology 284:228-243

20 Landis J, Koch G (1977) The measurement of observer agreement for categorical dataBiometrics, pp 16

21 National Lung Screening Trial Research T, Aberle DR, Berg CD et al (2011) The National Lung Screening Trial: overview and study design. Radiology 258:243-253

22 Picozzi G, Paci E, Lopez Pegna A et al (2005) Screening of lung cancer with low dose spiral CT: results of a three year pilot study and design of the randomised controlled trial "Italung-CT". Radiol Med 109:17-26

23 Wahidi MM, Govert JA, Goudar RK, Gould MK, McCrory DC, American College of Chest P (2007) Evidence for the treatment of patients with pulmonary nodules: when

is it lung cancer?: ACCP evidence-based clinical practice guidelines (2nd edition). Chest 132:94S-107S

24 Baldwin DR, Callister ME, Guideline Development G (2015) The British Thoracic Society guidelines on the investigation and management of pulmonary nodules. Thorax 70:794-798

25 Gierada DS, Pilgram TK, Ford M et al (2008) Lung cancer: interobserver agreement on interpretation of pulmonary findings at low-dose CT screening. Radiology 246:265-272

26 Ridge CA, Yildirim A, Boiselle PM et al (2016) Differentiating between Subsolid and Solid Pulmonary Nodules at CT: Inter- and Intraobserver Agreement between Experienced Thoracic Radiologists. Radiology 278:888-896

27 Martini K, Barth BK, Nguyen-Kim TD, Baumueller S, Alkadhi H, Frauenfelder T (2016) Evaluation of pulmonary nodules and infection on chest CT with radiation dose equivalent to chest radiography: Prospective intra-individual comparison study to standard dose CT. Eur J Radiol 85:360-365

28 Perry N, Broeders M, de Wolf C, Tornberg S, Holland R, von Karsa L (2008) European guidelines for quality assurance in breast cancer screening and diagnosis. Fourth edition--summary document. Ann Oncol 19:614-622

29 Revel MP, Bissery A, Bienvenu M, Aycard L, Lefort C, Frija G (2004) Are two-dimensional CT measurements of small noncalcified pulmonary nodules reliable? Radiology 231:453-458

30 Bogot NR, Kazerooni EA, Kelly AM, Quint LE, Desjardins B, Nan B (2005) Interobserver and intraobserver variability in the assessment of pulmonary nodule size on CT using film and computer display methods. Acad Radiol 12:948-956

31 Henschke CI, Naidich DP, Yankelevitz DF et al (2001) Early lung cancer action project: initial findings on repeat screenings. Cancer 92:153-159

32 Takashima S, Sone S, Li F et al (2003) Small solitary pulmonary nodules (< or =1 cm) detected at population-based CT screening for lung cancer: Reliable high-resolution CT features of benign lesions. AJR Am J Roentgenol 180:955-964

33 Liu X, Hou F, Qin H, Hao A (2018) Multi-view multi-scale CNNs for lung nodule type classification from CT images. Pattern Recognition 77:262-275

34 Bankier AA, Kressel HY (2012) Through the Looking Glass revisited: the need for more meaning and less drama in the reporting of dose and dose reduction in CT. Radiology 265:4-8

Figure captions

Figure 1: Nodule classification as assessed by the six readers. The submillisievert CT scans were rated as either no nodule present = N1; clearly benign nodule containing fat (a) or calcification (b) = N2; any other nodule < 5 mm = N3; nodules between 5 – 10 mm (d) = I; or nodules > 10 mm (e, f) = P. These findings were further subdivided in “negative findings” (N1, N2, N3) where no further action is needed and indeterminate findings (I) as well as positive findings (P).

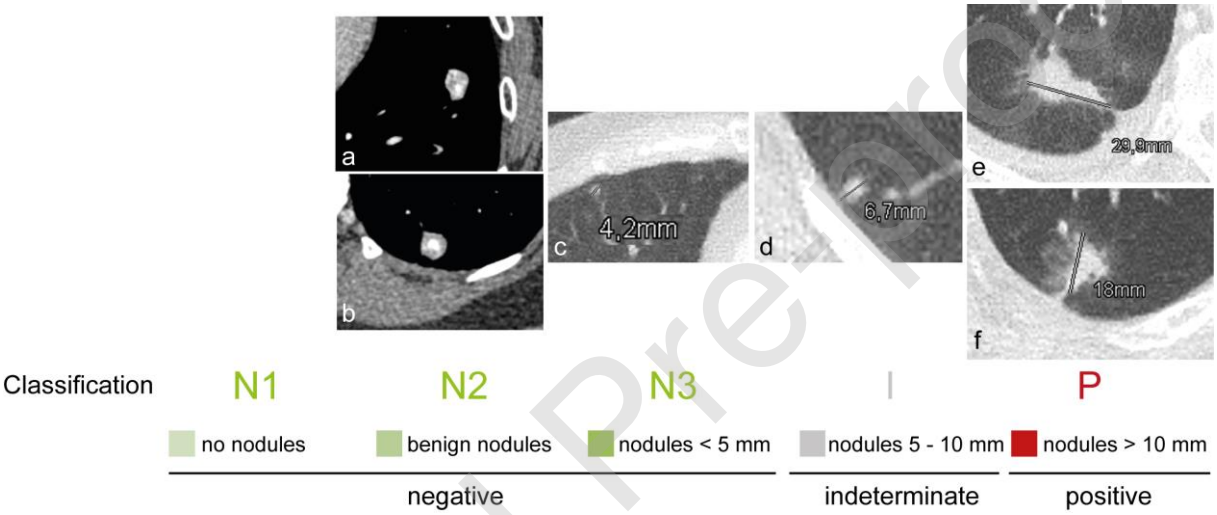
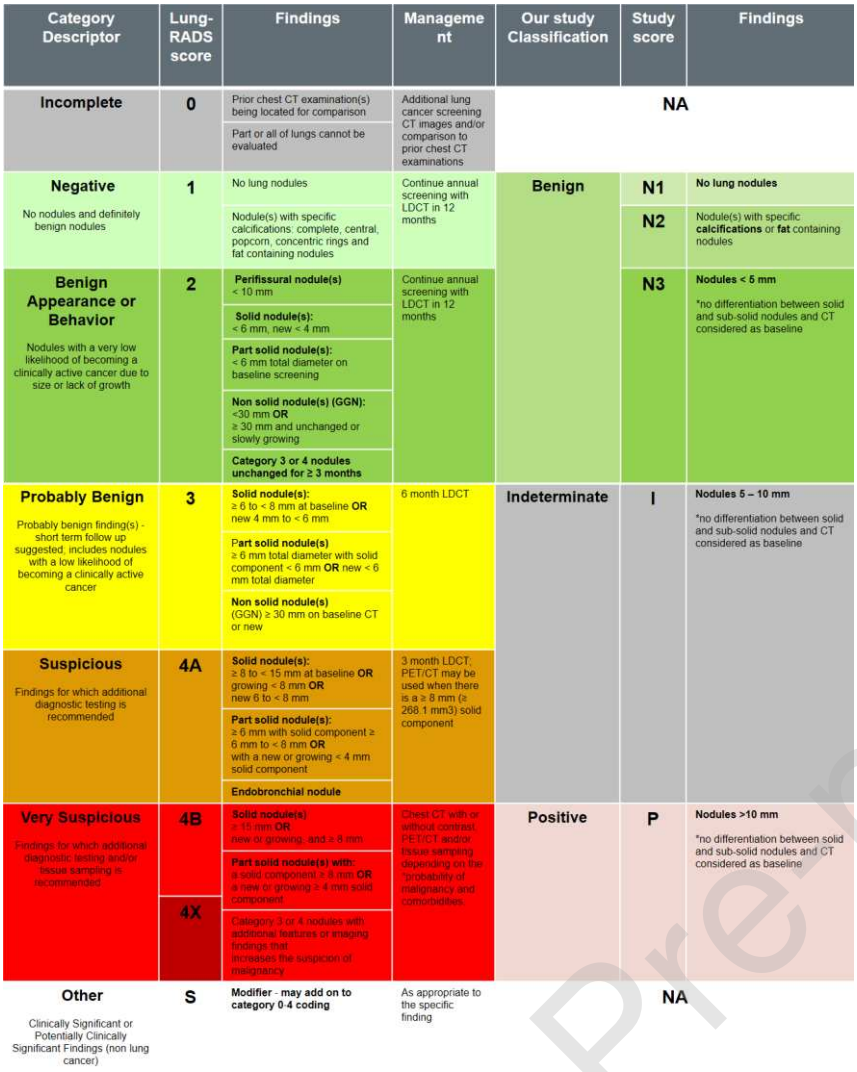


Figure 2: Our study classification system linked to the ACR/Lung-RADS system; Source: https://www.acr.org/Clinical-Resources/Reporting-and-Data-Systems/Lung-Rads



Category Descriptor	Lung-RADS score	Findings	Management	Our study Classification	Study score	Findings
Incomplete	**0**	Prior chest CT examination(s) being located for comparison Part or all of lungs cannot be evaluated	Additional lung cancer screening CT images and/or comparison to prior chest CT examinations	**NA**		
Negative No nodules and definitely benign nodules	**1**	No lung nodules	Continue annual screening with LDCT in 12 months	**Benign**	**N1**	**No lung nodules**
		Nodule(s) with specific calcifications: complete, central, popcorn, concentric rings and fat containing nodules			**N2**	Nodule(s) with specific **calcifications** or **fat** containing nodules
Benign Appearance or Behavior Nodules with a very low likelihood of becoming a clinically active cancer due to size or lack of growth	**2**	**Perifissural nodule(s)** < 10 mm **Solid nodule(s):** < 6 mm, new < 4 mm **Part solid nodule(s):** < 6 mm total diameter on baseline screening **Non solid nodule(s) (GGN):** <30 mm **OR** ≥ 30 mm and unchanged or slowly growing **Category 3 or 4 nodules unchanged for ≥ 3 months**	Continue annual screening with LDCT in 12 months		**N3**	**Nodules < 5 mm** *no differentiation between solid and sub-solid nodules and CT considered as baseline
Probably Benign Probably benign finding(s) - short term follow up suggested; includes nodules with a low likelihood of becoming a clinically active cancer	**3**	**Solid nodule(s):** ≥ 6 to < 8 mm at baseline **OR** new 4 mm to < 6 mm **Part solid nodule(s)** ≥ 6 mm total diameter with solid component < 6 mm **OR** new < 6 mm total diameter **Non solid nodule(s)** (GGN) ≥ 30 mm on baseline CT or new	6 month LDCT	**Indeterminate**	**I**	**Nodules 5 – 10 mm** *no differentiation between solid and sub-solid nodules and CT considered as baseline
Suspicious Findings for which additional diagnostic testing is recommended	**4A**	**Solid nodule(s):** ≥ 8 to < 15 mm at baseline **OR** growing < 8 mm **OR** new 6 to < 8 mm **Part solid nodule(s):** ≥ 6 mm with solid component ≥ 6 mm to < 8 mm **OR** with a new or growing < 4 mm solid component **Endobronchial nodule**	3 month LDCT; PET/CT may be used when there is a ≥ 8 mm (≥ 268.1 mm3) solid component			
Very Suspicious Findings for which additional diagnostic testing and/or tissue sampling is recommended	**4B** **4X**	**Solid nodule(s)** ≥ 15 mm **OR** new or growing, and ≥ 8 mm **Part solid nodule(s) with:** a solid component ≥ 8 mm **OR** a new or growing ≥ 4 mm solid component Category 3 or 4 nodules with additional features or imaging findings that increases the suspicion of malignancy	Chest CT with or without contrast, PET/CT, and/or tissue sampling depending on the *probability of malignancy and comorbidities.	**Positive**	**P**	**Nodules >10 mm** *no differentiation between solid and sub-solid nodules and CT considered as baseline
Other Clinically Significant or Potentially Clinically Significant Findings (non lung cancer)	**S**	**Modifier - may add on to category 0-4 coding**	As appropriate to the specific finding	**NA**		

Figure 3: Frequency distribution of different nodule classification among the six readers. Study CT scans were classified as: No nodule present, benign nodules present (e.g., containing calcium or fat), any other nodule measuring < 5 mm, any other nodule measuring between 5 – 10 mm (indeterminate) and any nodule measuring > 10 mm (positive).

Number of years of experience (y).

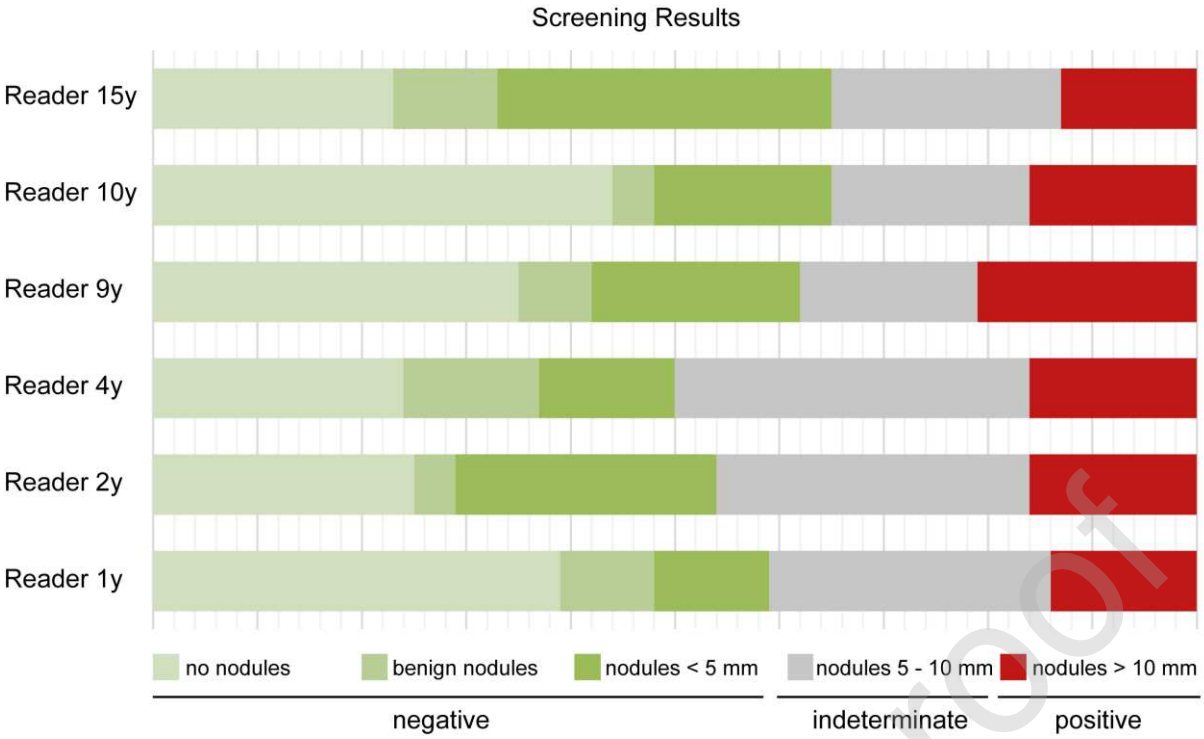


Figure 4: Reader results illustrating positive vs. non positive ratings (A) as well as cases as rated as either positive or intermediate vs. negative (B).

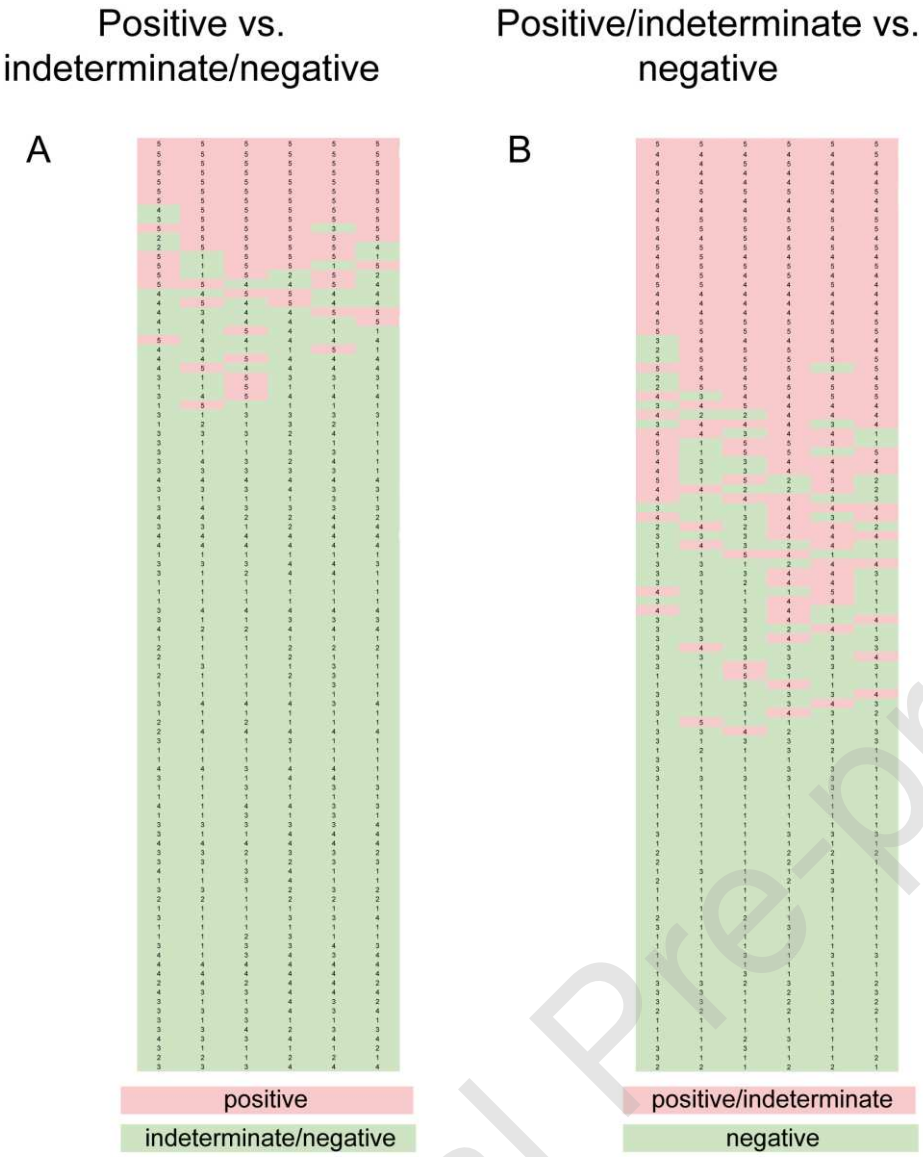
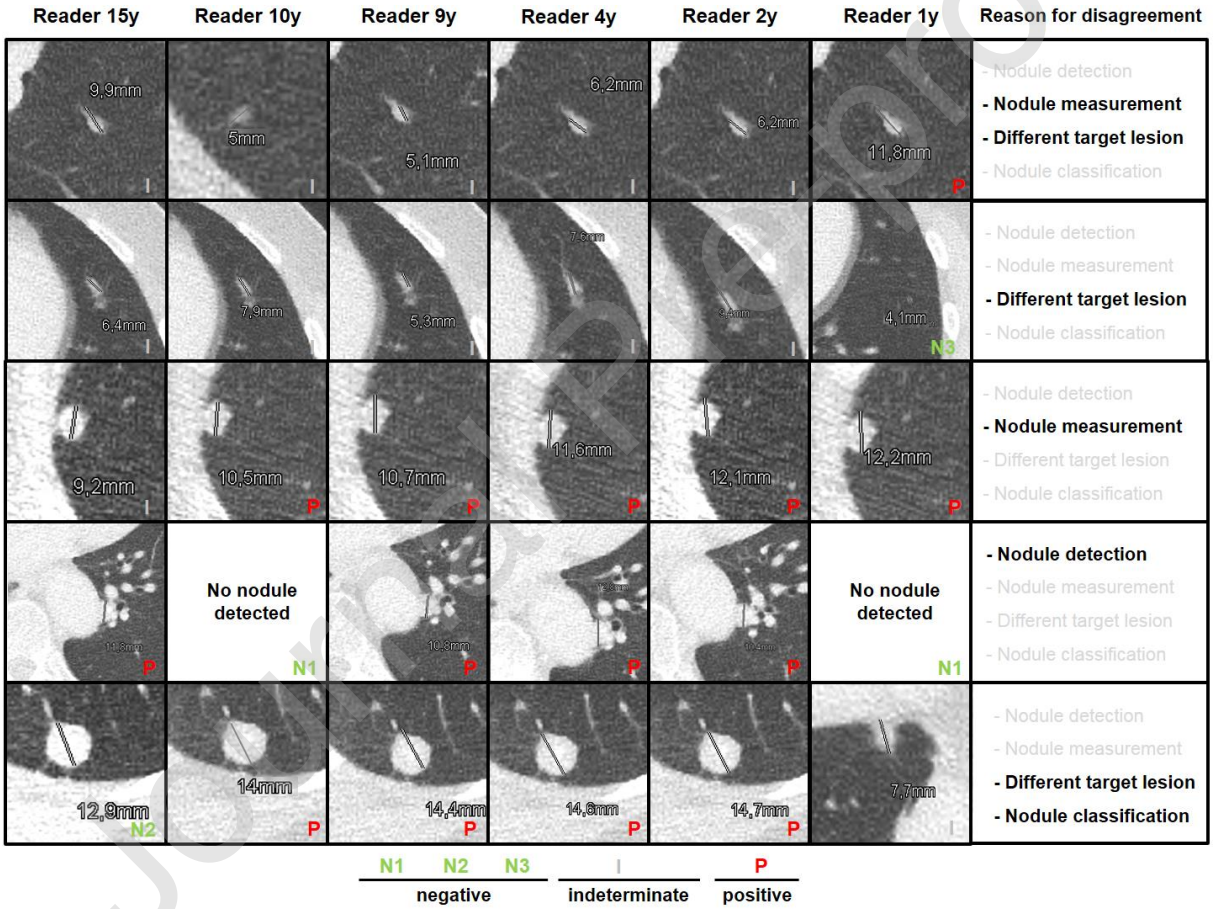


Figure 5: Illustration of the most frequent reasons of disagreement among the six readers including nodule detection, nodule measurement, identification of different target lesion, and nodule classification. In row 1, five out of six readers rated the scan as indeterminate (I), while the radiologist with 1y of experience rated the scan as Positive (P) due to different nodule measurement (range 5 - 12 mm). The reader with 10y of experience has chosen a different target leasion, however this did not lead to a different classification ot the CT. In row 2, the reader with 1y of experience has chosen a different target lesion which led to different classification ot the CT (N3 vs. I). In row 3, all readers have identified the same target lesion, however, there was different

classification of the CT due to different nodule measurement (range 9 - 12 mm). In row 4, four out of six readers rated the scan as Positive (P) due to a nodule > 10 mm adjacant to the descending aorta, while the remaining two readers rated the scan as Negative (N). This was due to missed nodule detection, potentially due to its close position to the descending aorta. In row 5, four out of six readers rated the scan as Positive (P) due to a nodule > 10 mm in the left lower lobe, while the reader with 15y of experience rated the scan as N2 and suggested the nodule was calcified. The reader with 1y of experience has chosen a different target lesion which led to different classification ot the CT the other readers.



Number of years of experience (y).

Figure 6: Correlation of reading time and reader experience.

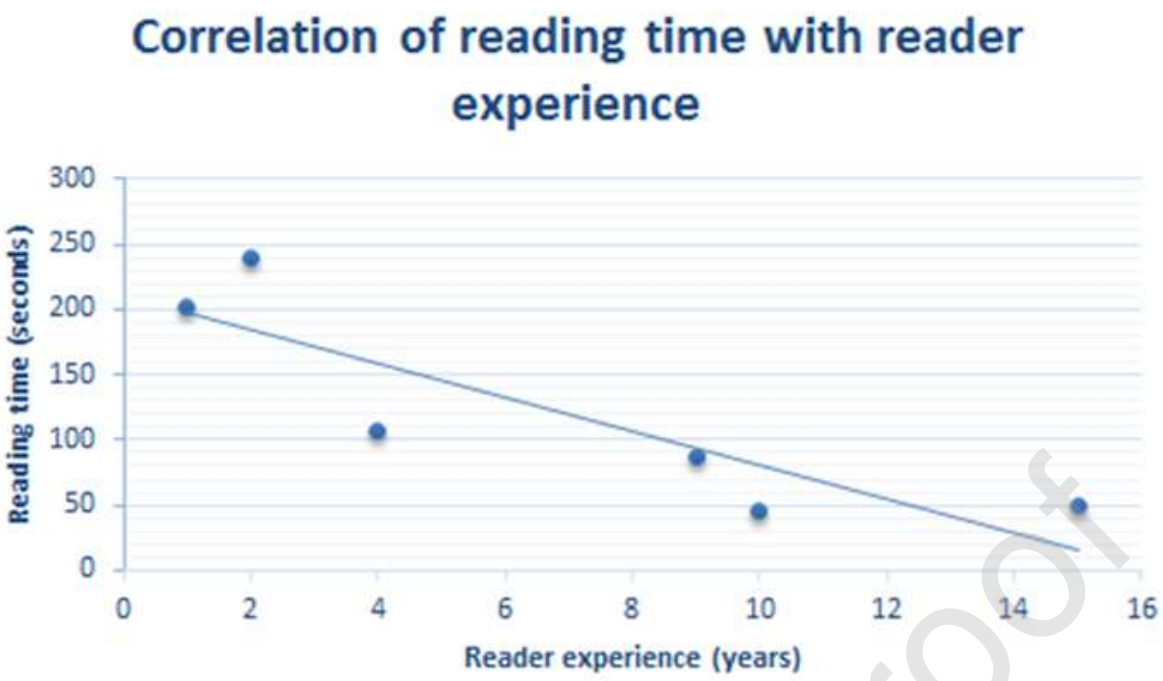


Table 1 Nodule classification as assessed by the six readers.

Category	**Features**	**Screening result**	**Code**
No nodule		Negative	N1
Benign nodules	Calcification Fat	Negative	N2/C N2/F
Nodules < 5 mm		Negative	N3
Nodules 5 - 10 mm		Indeterminate	I
Nodules > 10 mm		Positive	P

Table 2 Interreader agreement for nodule classification.

Reader Agreement	**Kappa**	***p*-value**
Overall	0.454	0.001
Experienced[a]	0.404	0.001
Unexperienced[b]	0.544	0.001
Reader 15y vs.		
Reader 1y	0.448	0.001
Reader 2y	0.484	0.001
Reader 4y	0.513	0.001
Reader 9y	0.401	0.001
Reader 10y	0.360	0.001
Reader 10y vs.		
Reader 1y	0.447	0.001
Reader 2y	0.316	0.001
Reader 4y	0.377	0.001
Reader 9y	0.450	0.001
Reader 9y vs.		
Reader 1y	0.440	0.001
Reader 2y	0.513	0.001
Reader 4y	0.423	0.001
Reader 4y vs.		
Reader 1y	0.543	0.001
Reader 2y	0.540	0.001
Reader 2y vs.		
Reader 1y	0.550	0.001

Number of years of experience (y), versus (vs.)

[a] i.e., readers with 9, 10, and 15 years of experience

[b] i.e., readers with 1, 2, and 4 years of experience

Table 3 Absolute nodule classification per reader. No nodule present (N1), nodule of any size with calcium content (N2/C), nodule of any size with fat content (N2/F), any other nodule measuring ≤ 4 mm (N3), any other nodule measuring ≥ 5 mm and ≤ 10 mm (I) and any nodule measuring > 10 mm (P).

	Reader 15y	**Reader 10y**	**Reader 9y**	**Reader 4y**	**Reader 2y**	**Reader 1y**
N1	23	44	35	24	25	39
N2/C	8	4	7	13	4	9
N2/F	2	0	0	0	0	0
N3	32	17	20	13	25	11
I	22	19	17	34	30	27
P	13	16	21	16	16	14

Number of years of experience (y), versus (vs)

Table 4 Differences in nodule measurement by groups.

Nodule size	**Mean, mm**	**SD**	**Mean range, mm**
Nodules < 5 mm	3.4	1.0	2.4
Nodules 5 - 10 mm	7.1	2.6	5.2
Nodules > 10 mm	19.7	9.2	6.4

SD standard deviation

Table 5 Reasons for disagreement among all six readers in the study patients (*n* = 100)

Reason for disagreement	**Number of cases overall** *n* = 155	**Number of cases were disagreement led to different management**[a] *n* = 77
Nodule detection	45 (29%)	21 (27%)
Nodule measurement	44 (28%)	23 (30%)
Different target lesion	40 (26%)	21 (27%)
Nodule classification	26 (17%)[b]	12 (16%)[c]

Data are presented as n (%)

[a] i.e., leading to a shift from a positive/indeterminate to a negative screening results and vice-versa

[b] including identification of calcification (*n* = 24) and fat (*n* = 2)

[c] including identification of calcification (*n* = 10) and fat (*n* = 2)